Received: 13 February 2018 | Revised: 25 May 2018 | Accepted: 15 July 2018

DOI: 10.1002/hbm.24333

WILEY

RESEARCH ARTICLE

Singing in the brain: Neural representation of music and voice as revealed by fMRI

Jocelyne C. Whitehead[1,2,3] | Jorge L. Armony[1,2,4]

[1]Douglas Mental Health University Institute, Verdun, Canada

[2]BRAMS Laboratory, Centre for Research on Brain, Language and Music, Montreal, Canada

[3]Integrated Program in Neuroscience, McGill University, Montreal, Canada

[4]Department of Psychiatry, McGill University, Montreal, Canada

Correspondence
Jocelyne C. Whitehead, Douglas Mental Health University Institute, 6875 LaSalle boulevard, Verdun, QC H4H 1R3, Canada.
Email: jocelyne.whitehead@mail.mcgill.ca

Funding information
Canadian Institutes of Health Research, Grant/Award Numbers: CIHR, MOP-130516, MOP-97967; Centre for Research on Brain, Language and Music, Grant/Award Number: CRBLM Graduate Student Stipend; Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: Canadian Graduate Scholarship- Masters fellowship NSERC, NSERC-CREATE Aud. Cog. Neurosci. Grad. Fellowship; 262439-2009, 2017-05832

Abstract

The ubiquity of music across cultures as a means of emotional expression, and its proposed evolutionary relation to speech, motivated researchers to attempt a characterization of its neural representation. Several neuroimaging studies have reported that specific regions in the anterior temporal lobe respond more strongly to music than to other auditory stimuli, including spoken voice. Nonetheless, because most studies have employed instrumental music, which has important acoustic distinctions from human voice, questions still exist as to the specificity of the observed "music-preferred" areas. Here, we sought to address this issue by testing 24 healthy young adults with fast, high-resolution fMRI, to record neural responses to a large and varied set of musical stimuli, which, critically, included *a capella* singing, as well as purely instrumental excerpts. Our results confirmed that music; vocal or instrumental, preferentially engaged regions in the superior STG, particularly in the anterior planum polare, bilaterally. In contrast, human voice, either spoken or sung, activated more strongly a large area along the superior temporal sulcus. Findings were consistent between univariate and multivariate analyses, as well as with the use of a "silent" sparse acquisition sequence that minimizes any potential influence of scanner noise on the resulting activations. Activity in music-preferred regions could not be accounted for by any basic acoustic parameter tested, suggesting these areas integrate, likely in a nonlinear fashion, a combination of acoustic attributes that, together, result in the perceived musicality of the stimuli, consistent with proposed hierarchical processing of complex auditory information within the temporal lobes.

KEYWORDS

fMRI, music; speech; singing, neural overlap, neural preference, pulse clarity

1 | INTRODUCTION

The syntactic parallels that music has with speech and its comparable use for communicating emotional states have contributed to a long-standing debate over a possible common evolutionary origin (Besson & Schön, 2001). Studies highlighting their similarities, at behavioral and neural levels, have encouraged the development of several theories attempting to make sense of the close relationship that music has to speech (for a recent review, see Peretz, Vuvan, Lagrois, & Armony, 2015). For example, Brown (2000) proposed the "musilanguage" hypothesis, stating that music and language have evolved from the same origin and over time diverged, adopting their own unique domain-specific attributes. Others hypothesized an invasion of music into the language module (i.e., a "functionally specialized cognitive system"; Fodor, 1983) now acting as an adapted by-product (Pinker, 1997) that has since stabilized across cultures (Sperber & Hirschfield, 2004). In contrast, others argue that the similarities between music and speech are not unique, as they are also shared with other cognitive mechanisms (Jackendoff, 2009). Attempts at reconciling these opposing views propose that music and language processing occur across a number of discrete modules, some of which overlap, while others remain distinct (e.g., Peretz & Coltheart, 2003).

The surge in neuroimaging studies conducted over the last decade that examined the neural correlates of speech and music processing, has rekindled this debate, particularly focusing the question on whether speech and music activate distinct or overlapping regions

in the brain, especially within the auditory cortex. As shown by a recent meta-analysis (Schirmer, Fox, & Grandjean, 2012), these studies have provided substantial evidence for overlapping regions of activation, in response to both music and voice, within the superior temporal gyrus (STG), superior temporal sulcus (STS), and medial temporal gyrus (MTG). However, a small but growing number of experiments, some using newly developed, more sensitive acquisition and/or analytical approaches, have reported some degree of functional separability of responses, with voice (including, but not limited, to speech) engaging mainly an area along the STS (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Belin, Zatorre, & Ahad, 2002; Belin & Zatorre, 2003; Fecteau, Armony, Joanette, & Belin, 2004; Kriegstein & Giraud, 2004; Pernet et al., 2015), and music eliciting stronger responses in a smaller cluster in the anterior STG (planum polare), often bilaterally, but more pronounced on the right hemisphere (Leaver & Rauschecker, 2010; Fedorenko, McDermott, Norman-Haignere, & Kanwisher, 2012; Angulo-Perkins et al., 2014; Aubé, Angulo-Perkins, Peretz, Concha, & Armony, 2015). Importantly, these findings were obtained with a variety of stimuli (e.g., Music: unfamiliar pop/rock music, instrumental excerpts of piano, strings, woodwind, or brass; Voice: syllables, words, connected speech, nonlinguistic vocalizations, varying languages; Controls: scrambled music or voice, songbirds, animal sounds, nonvocal human sounds, white noise, environmental, and mechanical sounds) and paradigms (e.g., block and event-related designs). Moreover, these results obtained using category-based univariate analyses were confirmed by a few others employing data-driven classification techniques based on multivariate statistics (Norman-Haignere, Kanwisher, & McDermott, 2015; Rogalsky, Rong, Saberi, & Hickok, 2011), as well as adaptation fMRI (Armony, Aubé, Angulo-Perkins, Peretz, & Concha, 2015).

Although most of the studies previously described attempted to control for the possible nonspecific effects of general acoustic characteristics of the stimuli employed (e.g., duration, intensity, and frequency), there are still important qualitative and quantitative differences between instrumental music and voice, which could, in principle, introduce confounds in the results obtained.

While it is impossible, and indeed undesirable (Leaver & Rauschecker, 2010), to remove all possible acoustic differences between music and speech (the same way it is not possible to do so for vocal vs. nonvocal sounds, or face to nonface visual stimuli), it is important to minimize them, leaving only those features that are thought to be essential to each stimulus class. In this sense, lyrical song as produced by the human voice in the absence of instruments, or *a capella*, may constitute an ideal candidate as an intermediary between music and speech (Schön et al. (2010)). Indeed, while singing is undoubtedly a form of musical expression, its basic acoustic profile is highly similar to that of the spoken voice. In fact, a "super-expressive voice" theory of music has been put forward, suggesting that music originated simply as an exaggeration of speech, accentuating vocal speed, intensity, and timbre, as a method of enhancing communication and to ensure effective bonding (Juslin, 2001). The few studies that directly compared brain responses to speech and singing support, to some extent, this hypothesis. Schön et al. (2010) presented subjects with French tri-syllabic nouns either spoken or sung, and observed that both conditions activated, as compared with pink noise, similar clusters in the middle and superior temporal gyrus bilaterally. The comparison of Singing > Speech revealed only small clusters in those regions, leading the authors to conclude that very similar networks are engaged when listening to spoken or sung words. Callan et al. (2006) compared six well-known songs in spoken and sung form and also found very similar activation patterns for both categories. They also reported greater activity for the singing than speech condition in the right planum temporale. However, these studies did not include an instrumental music condition, so the question remains as to whether there are brain regions that respond preferentially to music, regardless of how it is expressed, either through voice or instruments.

Another potential concern when conducting fMRI studies using acoustic stimuli is the possible influence of scanner noise in the observed responses. Although a large literature exists consistently showing that auditory perception studies can be successfully conducted using standard continuous acquisition sequences, it is still generally acknowledged that the use of sparse sampling protocols, or "silent fMRI"—in which the sounds are presented during a silent period, with volume acquisitions following the silence when the hemodynamic response function is at its peak (Hall et al., 2014)—does present advantages (as well as drawbacks, particularly in terms of reduced statistical power; Nebel et al., 2005). For instance, studies that have compared the two approaches have shown the recruitment of larger networks using sparse sampling (Adank, 2012), as well as greater activation in auditory regions (Gaab, Gabrieli, & Glover, 2006), and a higher MR signal-to-noise ratio (Hall et al., 1999). Furthermore, it has been suggested that speech perception in the presence of background noise requires the recruitment of additional cognitive resources as to successfully understand what is being spoken (Manan, Yusoff, Franz, & Mukari, 2013) and that it can impair other cognitive processes, such as memory recall (Rabbit, 1968; Murphy, Craik, Li, & Schneider, 2000). Moreover, noisy speech has been shown to elicit stronger responses in several brain regions, including middle and superior temporal gyrus (Davis & Johnsrude, 2003). Because the majority of fMRI studies of music perception employed continuous acquisition, it remains unknown to what extent, if at all, scanner noise may have affected the results obtained.

The goal of the present study was thus to provide a comprehensive assessment of the brain responses to music, including both instrumental and vocal (singing) stimuli. We employed a large and diverse set of unfamiliar short stimuli and controlled, either in the stimulus selection or analysis, many of the basic acoustic parameters. Analysis was conducted using complementary uni- and multivariate approaches. We employed a multiband echo-planar imaging sequence, in which the acceleration of data acquisition allowed us to achieve both high spatial resolution and sampling rate (thus maximizing statistical power). In addition, we conducted, in the same subjects, a short experiment using a subset of the stimuli using the *Interleaved Silent Steady State (ISSS)* sparse imaging acquisition protocol (Schwarzbauer, Davis, Rodd, & Johnsrude, 2006), to investigate the possible confounding effects of scanner noise on the results.

We expected to replicate previous studies showing that instrumental music, when compared with speech, activates a bilateral region in the anterior STG, particularly in the planum polare (PP) (Armony

et al., 2015; Leaver & Rauschecker, 2010; Angulo-Perkins et al., 2014; Fedorenko et al., 2012; Rogalsky et al., 2011), whereas speech would elicit responses along the STS (Belin et al., 2000, 2002; Belin & Zatorre, 2003; Fecteau et al., 2004; Kriegstein & Giraud, 2004; Pernet et al., 2015). Critically, we hypothesized that vocal music (i.e., singing) would represent an intermediate condition between these two. Namely, when compared with music, singing should activate STS, but when compared with speech, it should yield activations overlapping with those associated with instrumental music within the PP.

2 | METHODS

2.1 | Participants and procedure

Twenty-four healthy volunteers (11 females, mean age = 25.5) with a range of musical expertise (years of training: *M* = 4.2, *SD* = 4.73) participated in the study. Participants had normal hearing and were right-handed. All subjects were fluent in English. Eleven of them also spoke another language, and nine spoke three languages. Overall, languages understood by the participants included English (24), Finish, French (11), German, Greek, Hindi, Italian (7), Malayalam, Mandarin (3), Persian, Spanish, and Turkish.

The experiment consisted of three 8-min runs, two using a continuous multi-band sequence and one with an interleaved silent steady state (ISSS) sequence (Schwarzbauer et al., 2006), described below. Participants passively listened to auditory stimuli while watching nature scenes. Stimuli were presented using E-Prime 2.0 (Psychology Software Tools) and delivered binaurally from MRI-compatible headphones (Model S14, Sensimetrics). A sound test was conducted prior to each testing session to confirm that the acoustic stimuli were audible in the presence of the background scanner noise for the continuous acquisitions and not too loud for the sparse sampling one. Functional images were acquired on a 3T Siemens TIM TRIO MRI scanner with a 32-channel head coil. In addition to the functional runs, a high-resolution 3D T_1-weighted image (voxel size = $1 \times 1 \times 1$ mm^3) was acquired using a magnetization-prepared rapid acquisition gradient echo (MPRAGE) sequence (TR = 2.3 s; TE = 3 ms, 192 slices) for anatomical co-registration and normalization.

2.2 | Stimuli

Auditory stimuli belonged to three categories:

2.2.1 | Instrumental music

An assortment of instrumental pieces were cut to produce 60 different musical excerpts (duration: *M* = 1.49 s; *SD* = 0.13 s). The clips consisted of strings, woodwinds, or percussion instruments (40 unique instruments), each obtained from online database sources and from Vieillard et al. (2008) and Aubé et al. (2015).

2.2.2 | Speech

A total of 60 different phrases spoken in 45 languages (ranging from English, Spanish, and French to Baatonum, Gujarati, Mongolian, and Yiddish) and one stimulus with no words ("baby talk") produced by speakers including children (*n* = 2) and adults (33 male), were obtained from various online databases (duration: *M* = 1.51 s; *SD* = 0.22 s).

2.2.3 | Singing

Stimuli consisted of 60 different singing excerpt (duration: *M* = 1.51 s; *SD* = 0.23 s), sung by one or several individuals of varying ages, including male (*n* = 28) and female (*n* = 32), without instrumental accompaniment ("a cappella"), sung in 19 different languages (e.g., English, German, Arabic, Ilocano, Doabi, Hebrew) or without words (*n* = 6), including song excerpts produced by amateur and professional singers, lullabies, and religious chanting (e.g., Church choir and Torah reading). About 61% of these were monophonic, 37% homophonic, and 2% polyphonic.

All stimuli were monaural, but presented binaurally. The sounds were resampled to 32 bits, at a sample rate of 44,100 Hz, and adjusted for loudness by normalizing to the short-term loudness (STL) maximum using the Moore and Glasberg Loudness model (Glasberg & Moore, 2002), as implemented in the Loudness Toolbox on MATLAB. Basic acoustic parameters for each of the categories, computed using the MIRtoolbox (Lartillot, Toiviainen, & Eeorla, 2008), MATLAB scripts (Ewender, Hoffmann, & Pfister, 2009) and the Praat Vocal Toolkit (Boersma, 2002), are summarized in Table 1.

2.3 | fMRI acquisition and analysis

2.3.1 | Continuous acquisition

Each run consisted of 90 stimuli; 30 speech, 30 singing, and 30 instrumental music excerpts, which were presented in a pseudo-random fully balanced order (equal number of first-order transitions between categories), to remove any possible carry-over effects. Each stimulus was presented only once and the stimulus subsets used in each run were counterbalanced across subjects. The auditory stimuli were presented in a continuous design and were jittered using a brief ISI (duration: *M* = 2.49 s, *SD* = 0.20 s).

Functional images were acquired using a multiband accelerated pulse sequence with a factor of 12 (Setsompop et al., 2012). Eight hundred volumes (72 slices per volume, interleaved acquisition; FOV = 208 × 208 mm^2, matrix = 104 × 104, voxel size = 2 × 2 × 2 mm^3; TR = 529 ms; TE = 35 ms) were acquired. The first 10 scans of the run were discarded due to T1 saturation. Image pre-processing was conducted using SPM12 (Wellcome Department of Imaging Neuroscience, London, UK; http://www.fil.ion.ucl.ac.uk/spm). Functional images were spatially realigned to the first volume and normalized to the MNI152 template. The images were then smoothed using a 6 mm FWHM isotropic Gaussian kernel.

2.3.2 | Univariate analysis

Statistical analysis was performed for each subject using a univariate general linear model (GLM) in which the categories of interest (Instrumental Music, Singing, and Speech) were entered as boxcars of length equal to the stimulus duration, convolved with the canonical hemodynamic response function. Subject-specific contrast instrumental music versus speech, instrumental music versus singing, and singing versus speech, were then taken to a second level, repeated-measures ANOVA. Statistical significance was determined using a

TABLE 1 Mean and standard deviation values of acoustic features for each sound category

Audio features	Music	Singing	Speech
Articulation (a.u.)	.32 (.21)[a]	.27 (.15)[a]	.44 (.09)[a]
Root mean square (dB)	.13 (.05)	.16 (.04)[b]	.13 (.04)
Tempo (bpm)	125 (30)	137 (29)	126 (30)
Spectral centroid (kHz)	2.3 (1.5)	2.4 (1.0)	2.1 (1.0)
Spectral brightness (>1.5 kHz)	.44 (.26)	.42 (.16)	.37 (.15)
Spectral spread (Hz)	5.8 (3.3)	6.0 (2.6)	5.1 (1.6)
Spectral Skewness (a.u.)	.21 (.20)*	.37 (.36)*	.30 (.28)
Spectral kurtosis (a.u)	.73 (1.09)	1.5 (3.0)	.86 (1.43)
Spectral roll off 95th percentile (kHz)	4.2 (2.7)	5.0 (2.3)	4.3 (2.0)
Spectral Spectentropy (bits)	.76 (.08)[b]	.80 (.05)	.81 (.04)
Spectral flatness	.05 (.08)	.06 (.05)*	.04 (.03)*
Spectral irregularity	.78 (.32)	.95 (.38)*	.67 (.37)*
Zerocross (s^{-1})	1,335 (1,206)	97(517)	1,137 (548)
Low energy ratio	.54 (.10)	.48 (.08)[b]	.52 (.07)
Key clarity (a.u.)	6.8 (3.3)	6.0 (3.3)	7.1 (3.4)
Tonal mode (minor-major, a.u.)	−.02 (.12)	−.02 (.10)	−.02 (.08)
Pulse clarity (a.u.)	.28 (.17)[b]	.18 (.09)	.23 (.08)
Mean fundamental frequency (F0)	275 (138)	273 (90)	185 (56)[b]
Std. Dev. Fundamental frequency (F0)	47.2 (37.9)[b]	31.6 (24.9)	29.6 (15.4)
Minimum fundamental frequency (F0)	204 (109)	217 (79)	134 (45)[b]
Maximum fundamental frequency (F0)	353 (169)	327 (111)	246 (78)[b]
Fraction of locally unvoiced frames (%)	10.6 (13.3)	7.6 (9.1)	23.7 (13.1)[b]
Jitter (local) (%)	2.24 (2.57)	1.43 (1.19)[b]	2.25 (.71)
Shimmer (local) (%)	12.7 (6.8)[b]	10.1 (5.1)	10.2 (3.2)
Mean HNR	11.4 (8.0)	13.7 (5.4)	11.4 (3.1)

a.u. = arbitrary units; bpm = beats per minute. Values were calculated with MIRToolbox, except for those related to the Fundamental Frequency (http://www.tik.ee.ethz.ch/~spr/f0_detection) and the last four features (Praat).
[a] All significantly different.
[b] Significantly different from the other two.
*Significantly different from each other ($p < .05$, Bonferroni corrected).

voxel threshold *of* $p = .001$, with a cluster-based familywise error rate (FWE) correction for multiple comparisons of $p < .05$ ($k = 90$) as implemented in AFNI's algorithm 3dClustSim (AFNI version 16.3.05). To identify regions commonly activated for different categories (e.g., Instruments and Singing vs. Speech), we performed conjunction analyses (minimum statistic compared with the conjunction null; Nichols, Brett, Andersson, Wager, & Poline, 2005).

In addition, we conducted a stimulus-based analysis. For each subject, each of the 180 stimuli was entered as a separate covariate in a standard GLM. The corresponding stimulus-specific parameter estimates were then averaged across subjects. These estimates were used for post-hoc regression analyses of the significant clusters including the acoustic parameters shown in Table 1, as well as for the multivariate analysis described in the following paragraph.

2.3.3 | Multivariate analysis

The categorical univariate analyses were complemented by a simple stimulus-based multivariate approach in which the parameter estimate images obtained in the stimulus-based analysis described in the previous paragraph were submitted to an Independent Component Analysis (ICA). We restricted the observations to auditory-responsive voxels as identified by an omnibus *F*-test in the univariate group analysis. Furthermore, Principal Component Analysis (PCA) was first applied on the data to reduce the dimensionality of the signal to the subspace spanned by the first four components, which explained 87% of the total variance. The contributions of each stimulus to each of the independent components obtained (weights) were then submitted to independent-sample *t*-tests (Bonferroni-corrected for multiple tests) to assess whether there were significant differences between categories (Speech, Instrumental Music, and Singing). Finally, the weights were submitted to a multiclass, *error-correcting output codes* (ECOC) model (a generalization of support-vector machine classification for more than two classes; Dietterich & Bakiri, 1995), implemented in MATLAB, to determine if the model could classify individual stimuli as belonging to their a priori category with above-chance accuracy.

2.4 | Sparse acquisition

Functional images were acquired using the ISSS sequence (Schwarzbauer et al., 2006) (FOV = 224 × 224 mm^2, matrix = 104 × 104, voxel size = 2 × 2 × 2 mm^3; TR = 2,383 ms; TE = 30 ms), with 25 slices, parallel to the Sylvian fissure, covering the entire of auditory cortex. Seven TRs formed a single epoch, in which three of the volumes were acquired during the silent dummy block (no data acquisition), followed by four volumes during the acquisition block. Auditory stimuli were presented during the silent periods in a short block of four stimuli belonging to the same category (Instrumental Music or Speech) with a mean duration of 7.15 s and their onset relative to the beginning of the dummy block was jittered (latency: 0.617 ± 0.403 s). Because of time limitations, due to the longer time required to acquired images, only two categories were presented, Instrumental Music and Speech. A total of 48 stimuli per category, taken from the stimulus pool described above, were presented in 12 blocks, their order within and between blocks, pseudo-randomized, and counterbalanced across participants. In addition, there were six blocks of silence, which served for baseline estimation.

Data preprocessing was carried out as in the continuous acquisition (see above) and analysis was performed using a finite impulse response (FIR) model, in the context of the general linear model, in which each of the four acquisition volumes for the two sound types was entered as a separate category (i.e., eight in total). Dummy volumes were created using replications of the mean EPI image, to create a continuous timescale in the design matrix. The dummy scans were not included as observations in the model, to avoid skewing the degrees of freedom (Peelle, 2014). Subject-specific estimates for the contrast for Instrumental Music minus Speech were calculated and taken to a second level, one-sample *t*-test. Statistical significance was determined as in the previous analysis. Analyses were also conducted using a hrf model, yielding similar results (not shown).

To compare the results between the continuous and sparse acquisitions, conjunction analyses were conducted for each contrast of interest (p = .01). Additionally, we tested whether there was a correlation in the magnitude of the responses between acquisitions, by entering the corresponding cluster-averaged, subject-specific contrast estimates into a linear regression analysis.

3 | RESULTS

3.1 | Continuous acquisition

3.1.1 | Univariate analysis

Coordinates, z-scores, and cluster extents for all the significant activations obtained in the univariate analysis are reported in Table 2. The contrasts Instruments minus Speech yielded significant clusters in the right planum temporale (PT) and bilaterally in the planum polare (PP) (Figure 1a). Singing minus Speech yielded significant clusters bilaterally in the PP and in the right PT. Importantly, these clusters partially overlapped with those obtained in the preceding contrast (Figure 1b). This common activation for musical stimuli in general was statistically confirmed through a conjunction analysis ([Instruments − Speech] and [Singing − Speech]), which yielded significant activations in the right PT and bilateral regions in the PP (Figure 1a,b). Interestingly, in the more anterior regions of PP, Singing elicited stronger responses than both Speech and Instruments, whereas in the more posterior areas activation for Instruments was larger than for Speech and Singing (Figure 1b). Moreover, responses in this latter cluster, particularly in the left hemisphere, significantly correlated with stimuli's pulse clarity values (z = 3.57, p < .001).

Speech vs. Instrumental Music revealed significant bilateral activity in in voice-preferred areas within the superior temporal sulcus (STS), superior temporal gyrus (STG), and medial temporal gyrus (MTG) (Figure 2, *Top*). Largely overlapping activations were obtained for the contrast Singing vs. Instruments, confirmed statistically using a conjunction analysis (Figure 2). Finally, the contrast Speech versus Singing also yielded significant clusters bilaterally in the STS, STG, and MTG (Table 2).

To assess whether the responses of voxels activated in the contrasts [Instruments − Speech] and [Singing − Speech] were modulated by simple acoustic parameters, we extracted the stimulus-specific parameter estimates for each of the three clusters reported in Table 1 and entered them (one at a time), as an additional covariate in the analysis. None of the acoustic features significantly correlated with the BOLD parameter estimates in the music-preferred clusters (but see above for a correlation with pulse clarity in a subcluster of the contrast Instruments minus Singing and Speech).

In order to evaluate the robustness of the group-level activations of Music (Instruments and Singing) versus Speech, we tested for the presence of significant clusters in these contrasts for each subject separately, using an anatomical mask corresponding to the planum polare for each hemisphere, obtained from Harvard–Oxford Probabilistic Anatomical Atlas, as in our previous study (Angulo-Perkins et al., 2014). For the contrast Instruments minus Speech, 88% and 75% of subjects had significant clusters on the right and left hemispheres, respectively, using a significance threshold of p = .01 (uncorrected), and 75% and 63% with a more stringent threshold of p = .001. The proportion of subjects with significant clusters for the contrast Singing minus Speech was 83% and 75% for p = .01, and 63% and 58% for p = .001, for the right and left hemispheres, respectively. Figure 3 shows a prevalence map of the voxels, across the whole brain, that showed significant activation at the single-subject level (p = .01), for these two contrasts. Consistent with the group analysis (Figure 1), the individual clusters associated with Singing were slightly more anterior than those for Instrumental music.

3.1.2 | Multivariate analysis

The first ICA component (Figure 4a) included almost all voxels in the mask, representing, as expected, the general auditory responses elicited by all stimuli. There was a significant difference effect of category on the associated weights (F[2,179] = 4.867, p = .009), reflecting a smaller activation for Instruments compared with Speech (p = .01, Bonferroni corrected) and Singing (p = .06, Bonferroni corrected), with no difference between the two vocal sounds (p > .9). These results are in agreement with those from the univariate analysis.

The second ICA exhibited a bipolar pattern, with positive and negative subcomponents that largely overlapped with the music- and voice-preferred areas, respectively, obtained in the univariate conjunction analyses (Figure 1). Moreover, a significant category effect was observed for the corresponding weights, with all categories significantly differing from each other (all p's < .001). Interestingly, the scatterplot of the weights for each stimulus (Figure 4b) showed almost no overlap between Instruments (positive values) and Speech (negative values), whereas singing fell in between the two, consistent with the shared activation pattern of this category with both instrumental music and spoken voice. This separation among categories was confirmed through a multiclass, ECOC model, which yielded an overall classification accuracy was 68% (leave-one-out cross-validation, chance level: 33%; p < .0001). Similar results were obtained when analyzing only Speech and Singing, confirming that the results were not due to simple acoustic differences between instruments and human voice.

3.2 | Sparse acquisition

The contrast Instrumental Music minus Speech yielded significant clusters bilaterally in the PP and the right PT. Significant clusters in the bilateral STS, STG, and MTG were obtained for the contrast Speech minus Instrumental Music (Figure 5 and Table 2). Importantly, no additional activation clusters were observed for either of the comparisons when using the "silent" sparse sampling protocol. The location of the clusters is very similar to what we found with the continuous acquisition in the same group of subjects. Furthermore, there was a significant correlation of the subject-specific, cluster-averaged parameter estimates for the contrast Instrumental Music minus Speech between both runs for each of the three main clusters (0.44 < r < .56, p's < .05).

Examination of the contrasts at single-subject level, using the same approach as described for the continuous acquisition, revealed that 79% and 88% of the subjects had significant clusters on the left

TABLE 2 Significant activations associated with contrasts of interest at the group level

Anatomical location	Left			Right			Z-score (peak voxel)	K_E
	x	*y*	*z*	*x*	*y*	*z*		
Continuous multiband sequence								
Instrumental music > speech								
STG (posterior)				66	−28	12	5.99	155
STG (anterior)				46	−6	−6	5.88	241
STG (anterior)	−48	−6	−4				5.37	162
Singing > speech								
STG (anterior)				50	4	−8	6.29	319
STG (posterior)				66	−26	10	5.74	213
STG (anterior)	−48	0	−6				7.43	286
[Instrumental music and singing] > speech								
STG (posterior)				66	−26	10	5.74	123
STG (anterior)				48	4	−8	5.74	154
STG (anterior)	−48	−6	−4				5.37	85
Speech > instrumental music								
STS/STG, MTG				64	−8	−4	12.63	1,570
STG/STS, MTG	−62	−12	0				15.14	2008
Singing > instrumental music								
STS/STG, MTG				62	−20	−2	11.04	1,440
STS/STG, MTG	−60	−12	2				12.31	1835
[Speech and singing] > instrumental music								
STS/STG, MTG				62	−20	−2	11.04	1,250
STS/STG, MTG	−60	−12	1				12.31	1,612
Speech > singing								
STS/STG, MTG				64	−8	−4	7.22	845
STS/STG, MTG	−60	−22	−2				9.36	1,239
Interleaved silent steady state (ISSS) sequence								
Instrumental music > speech								
STG (posterior)				48	−32	24	4.39	238
STG (anterior)				42	−12	−10	4.88	344
STG (anterior)	−48	−6	−0				4.79	253
Speech > instrumental music								
STS/STG, MTG				56	−28	−2	7.54	863
STG/STS, MTG	−64	−28	4				8.01	1,238

STG = superior temporal gyrus; STS = superior temporal sulcus; MTG = medial temporal gyrus.

and right hemispheres, respectively, using a significance threshold of $p = .01$ (uncorrected), and 58% and 75% with a threshold of $p = .001$.

4 | DISCUSSION

The aim of this study was to identify the brain responses to vocal and musical stimuli through the use of a high spatial- and temporal-resolution fMRI sequence. By using a stimulus set that varied widely in most of the basic acoustic measures and, critically, by including both vocal and instrumental musical excerpts, we were able to minimize potential confounding effects caused by differences in physical properties between categories. Moreover, to rule out any possible influence of scanner noise on the observed activations, we also employed a "silent" (i.e., sparse sampling) acquisition sequence with a subset of the original stimuli, in the same subjects. Finally, results obtained with a univariate categorical analysis were confirmed by a stimulus-based, multivariate approach.

4.1 | Cortical responses to voice

When compared with musical instruments, human voice, either spoken or sung, elicited significant activations in clusters along the STS in both hemispheres. These results confirm and extend many reports in the literature showing that this region preferentially responds to the human voice (Belin et al., 2000, 2002; Belin & Zatorre, 2003; Fecteau et al., 2004; Kriegstein & Giraud, 2004; Pernet et al., 2015), with and without linguistic content. As the vocal stimuli included speech and

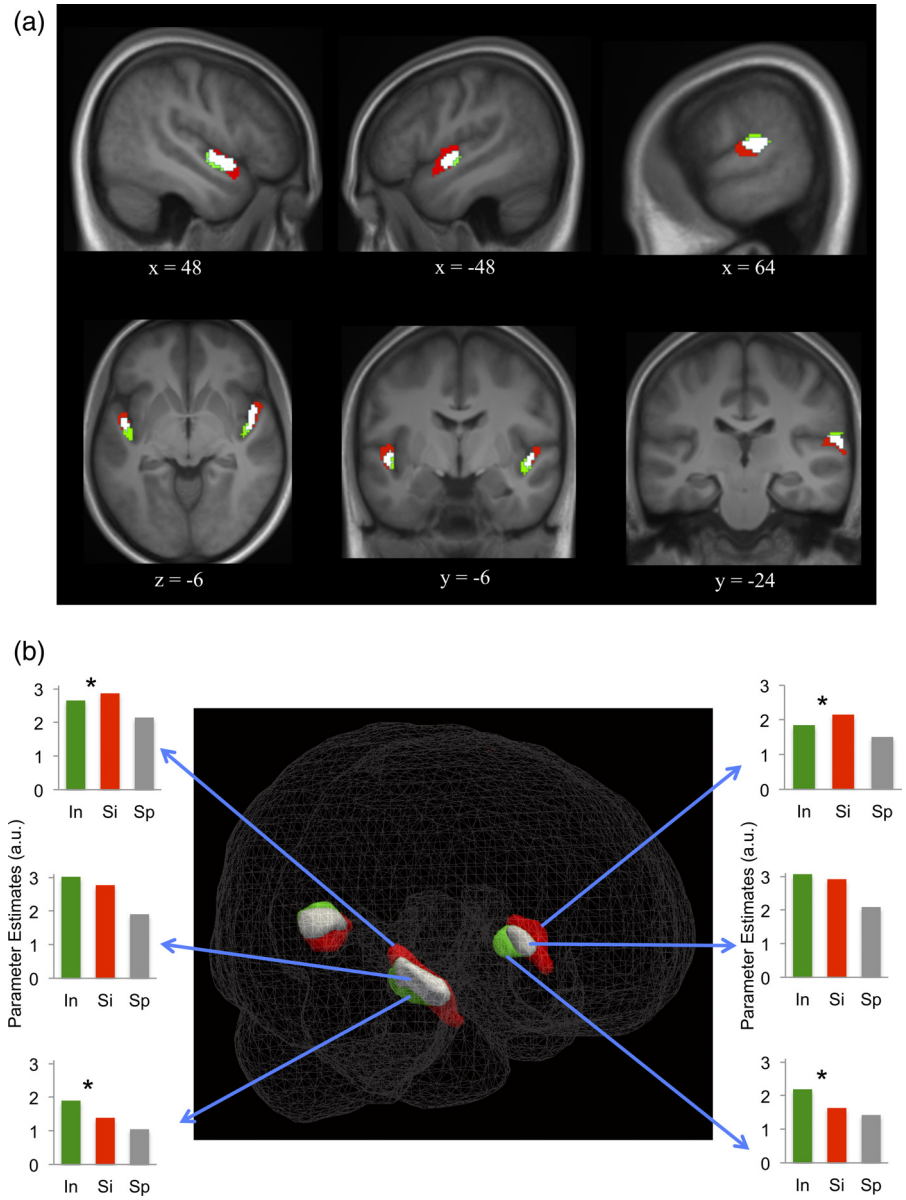

FIGURE 1 (a) 2D and (b) 3D renderings of the clusters of significant activations for the contrasts [singing − speech] (red), [instrumental music − speech] (green), as well as their conjunction (white). Threshold: $p = .001$ (corrected for multiple comparisons at the cluster level). Group average of the responses for each condition in each cluster (left and right hemispheres), using unsmoothed data. In: Instrumental music; Si: Singing; Sp: Speech; A.U.: arbitrary units. *significant difference ($p < .001$) between singing and instrumental music. In all cases, singing and instrumental music elicited significantly larger responses than speech ($p < .001$) [Color figure can be viewed at wileyonlinelibrary.com]

songs in different languages (most of which were not understood by the participants) as well as no-words singing, the responses observed in this area are likely to be related to the acoustic properties of the human voice, rather than reflecting semantic processing. This is also consistent with previous studies showing that this region responds significantly to human nonlinguisitic vocalizations (Belin et al., 2000, 2002; Fecteau et al., 2004). However, in this and previous studies, speech always elicited the strongest response, in both hemispheres. Importantly, and as reported before, while these clusters exhibited a bias, in terms of magnitude, for human voice, they also responded to nonvocal sounds, confirming that the so-called vocal temporal area (VTA) should be considered as a "voice-preferring" rather than as a "voice-selective" region (Belin et al., 2000).

4.2 | Cortical responses to music

Conversely, contrasted to speech, music—either in instrumental or vocal form—yielded significant clusters in the anterior planum polare bilaterally and in the right planum temporale, in agreement with previous studies employing different stimulus sets and analyses approaches (Patterson, Uppenkamp, Johnsrude, & Griffiths, 2002; Leaver & Rauschecker, 2010; Angulo-Perkins et al., 2014; Fedorenko et al., 2012; Rogalsky et al., 2011; Norman-Haignere et al., 2015; Aubé et al., 2015). In these previous studies most, when not all, musical stimuli contained an instrumental component, thus leaving open the question of whether these regions encode instrumental timbre (Leaver & Rauschecker, 2010), or music in general, including singing.

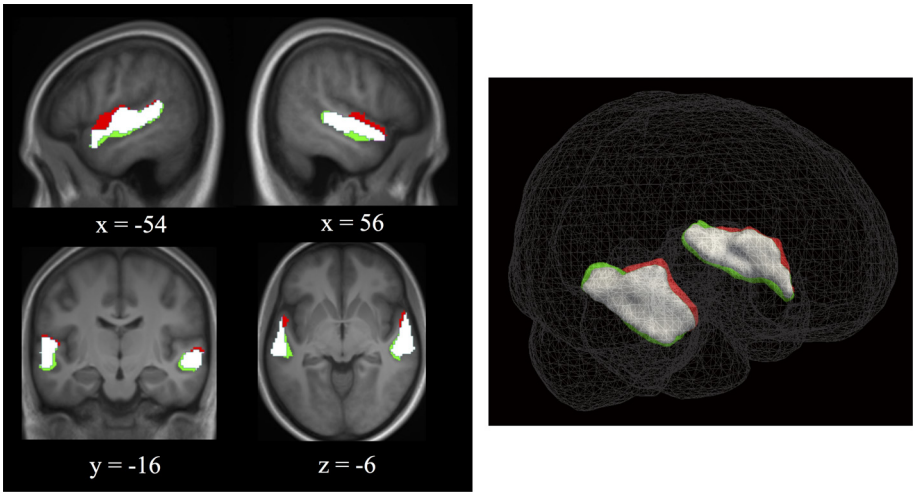

FIGURE 2 2D (*Left*) and 3D (*Right*) renderings of the clusters of significant activations for the contrasts [singing > instrumental music] (red), [speech > instrumental music] (green), as well as their conjunction (white). Threshold: $p = .001$ (corrected for multiple comparisons at the cluster level) [Color figure can be viewed at wileyonlinelibrary.com]

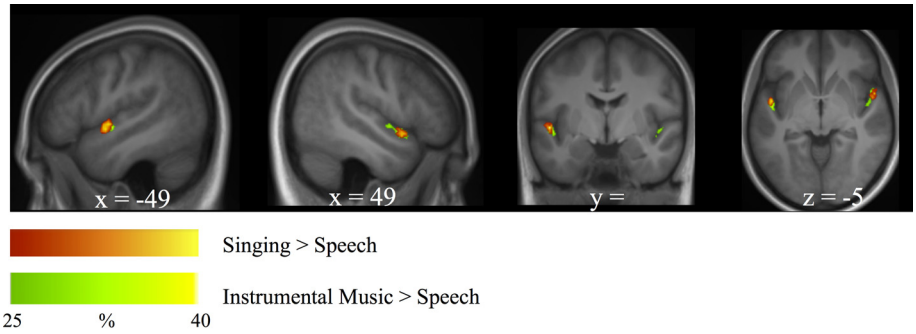

FIGURE 3 Prevalence maps showing the percentage of subject-specific significant activations at each voxel for the contrasts [singing > speech] (red scale) and [instrumental music > speech] (green scale). Clusters for singing were significantly more anterior (LH: $p = .008$; RH: $p = .02$) and lateral (LH: $p = .03$; RH: $p = .003$) than those for instrumental music [Color figure can be viewed at wileyonlinelibrary.com]

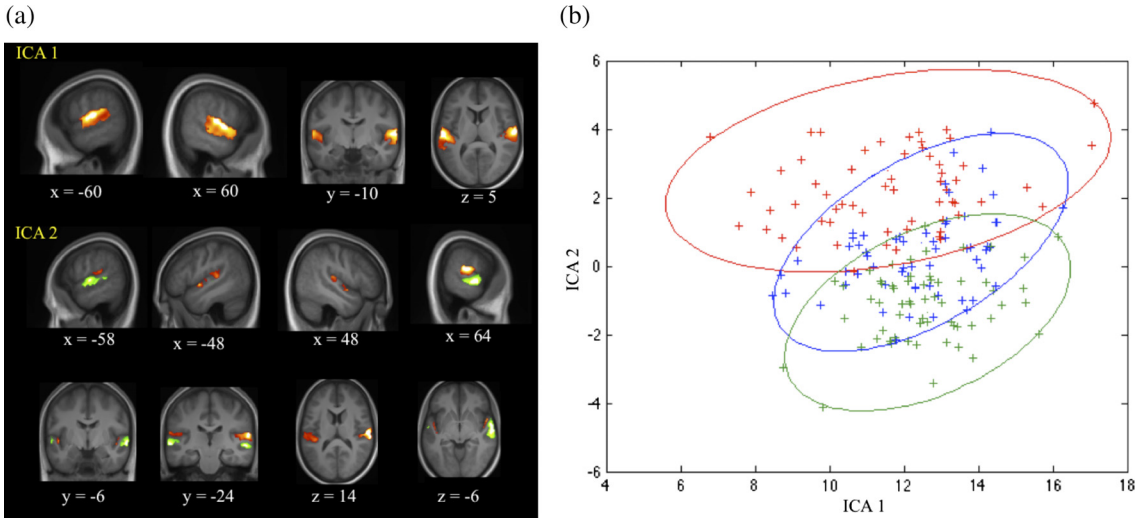

FIGURE 4 (a) First two components obtained in the stimulus-specific ICA. In the second component, red and green represent positive and negative values, respectively. (b) Scatterplots of the stimulus-specific eigenvalues corresponding to the first two ICA components. Each cross represents one stimulus: Instrumental music (red), singing (blue), and speech (green). Curves correspond to the minimum volume ellipsoid that covers all points of each category [Color figure can be viewed at wileyonlinelibrary.com]

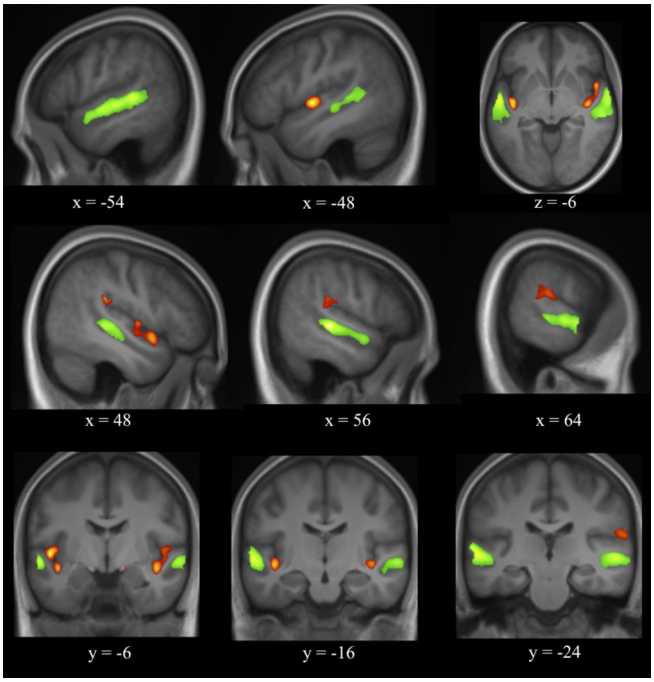

FIGURE 5 Clusters of significant activations for the contrasts [instrumental music > speech] (red) and [speech > instrumental music] (green) obtained with the sparse sampling acquisition. Threshold: p = .001 (corrected for multiple comparisons at the cluster level) [Color figure can be viewed at wileyonlinelibrary.com]

Our conjunction analysis directly answers this question, confirming that clusters within these regions respond more strongly to both instrumental and vocal music than speech, with no significant differences between the first two categories. Moreover, these results also help address the often-raised concern about the possible confounding effects due to differences in acoustic parameters between instrumental music and voice. Indeed, because of the acoustic similarities between spoken and sung vocal expressions, and the substantial differences between the latter and musical instruments (ranging from drums to guitars to xylophones), it is highly unlikely that these activations simply reflect differences in basic acoustic features among categories. Instead, these areas seem to encode a higher-order feature (yet, obviously, still based on the physical characteristics of the stimuli) that is shared among different forms of musical expression, more than with other complex, social stimuli such as speech (e.g., melody vs. sentence-level intonation; Zatorre & Baum, 2012). Such a conclusion is also supported by the lack of correlation of the responses with any of the tested basic acoustic parameters, also shown previously by Leaver and Rauschecker (2010). Finally, the almost identical results obtained with the sparse-sampling sequence, rules out potential differential effects of scanner noise on voice and music (see below for further discussion of these methodological issues).

As could be expected, the overlap of the instrumental and vocal music versus speech clusters was not complete. Specifically, more posterior regions of PP responded more strongly to instrumental music than both speech and singing. Interestingly, this cluster, particularly in the left hemisphere, significantly correlated with pulse clarity, in agreement with our previous study (Angulo-Perkins et al., 2014). This acoustic parameter measures the intrinsic rhythm of a stimulus, arguably one of the defining characteristics of (instrumental) music, and appears to be involved in musical genre recognition specifically. Pulse clarity improves the ability to discriminate between genres, which differ in how audible the main pulsation is, over the texture of the base rhythm (Lartillot, Eerola, Toiviainen, & Fornari, 2008). As the key organizing structure of music, rhythm is fundamental for melody and harmony to exist (Thaut, Trimarchi, & Parsons, 2014). In contrast, the more anterior portions of PP were activated significantly more to singing than to either speech or instruments. In this case, we failed to identify one, or a linear combination of, acoustic parameters that correlated with activity in this region, including factors previously identified as differentiating singing from speech, such as duration, fundamental frequency floor, and vocal intensity (Livingstone, Peck, & Russo, 2013). One possible explanation for this null result is that the transition from speech to song involves a more complex, nonlinear weighting of several acoustic features (Saitou, Tsuji, Unoki, & Akagi, 2004; Saitou, Goto, Unoki, & Akagi, 2007; Livingstone, Peck & Russo, 2013). Overall, the brief duration of the stimuli did not allow for the computation of additional information about the acoustic features to directly explore this question. Additionally, no effects were observed based on the number of voices or melody lines on the magnitude or location of the music-related activations, again, likely due to the small variability in these features present in our stimuli. Thus, longer stimuli, with discrete categorical differences as to properly analyze the acoustic attributes, may lead to a tangible conceptualization of music, and thus a worthwhile pursuit in future studies. Another, complementary approach could be to use stimuli that have been artificially manipulated, in the line of the work of Saitou et al. (2004, 2007) to obtain the necessary independent variability of these candidate parameters to attain the statistical power required for detecting small effects, and potentially shedding light on this question.

The notion that "music-preferred" respond to a complex configuration of varying acoustic components bears some parallels with observations made in the literature regarding the processing of visual social stimuli. For example, headless bodies have been shown to elicit a greater response in body-selective areas of the brain, when presented to participants as a whole configured body, rather than as separate segregated parts appearing together, but not in full form (Brandman & Yovel, 2016). It is most likely that processing musicality reflects this pattern, in which each acoustic component is required in a particular arrangement, as to induce this response. As suggested by the development of the speech-to-singing synthesis system (Saitou et al., 2004, 2007), it is also likely that varying weights of each acoustic modification must be precise for the musical perception to be achieved. This can be related to the observed saliency-hierarchy in the fusiform face area (FFA) in response to specific facial features. Lai, Pancaroglu, Oruc, Barton, and Davies-Thomson (2014)'s fMRI-adaptation study identified that different parts of the face (e.g., nose, mouth, and eyes) contribute varying amounts to the overall neural signal in face-sensitive regions of the brain, such that greater response sensitivity is present for the upper half of the face, and more specifically, the eyes. The origin, and specificity, of category-selective, or preferred, brain regions has also been extensively studied, and debated, in the visual domain. In particular, an alternative hypothesis to the view that face selectivity, of preference, in the FFA is hard-

wired (Kanwisher, McDermott, & Chun, 1997), has been put forward, suggesting that, rather than this region being face-sensitive, it may instead be better attributed as being an area of visual expertise, functioning to process and decipher highly complex visual stimuli (Gauthier, Skudlarski, Gore, & Anderson, 2000; Bilalić, 2016). According to this view, its preference for face stimuli reflects the fact that, as social individuals, we can all be considered experts of faces. Translating this idea to the auditory domain, it may be that as surrounded by music since birth, we have been able to fine-tune our perception of the particular "music algorithm," and now rely on this region of the brain to respond when needing to decipher more discrete changes within the musical framework. Some preliminary support for this idea comes from studies comparing responses to music between musicians and nonmusicians (Angulo-Perkins et al., 2014; Ohnishi et al., 2001), although further studies including musical expertise as a factor are still needed to fully test this hypothesis and better characterize the neural representation of musical and vocal stimuli in the brain.

4.3 | Methodological considerations

The location and extent of the clusters of significant activation in the contrasts Instrumental Music versus Speech (and vice versa) obtained with the continuous and sparse acquisitions were very similar, as shown in Figures 1–2, and 4. The concordance in results between the two sequences is in agreement with previous studies (e.g., Woods et al., 2009; Hall et al., 2014). Interestingly, we also found that the magnitudes of the responses for both runs were significantly correlated across subjects. Thus, it is very unlikely that differential effects of scanner noise on speech and music could have influenced the overall pattern of the observed activations. In turn, this provides further support for the use of continuous sampling sequences to study processing of complex auditory information, particularly when focusing on regions outside primary auditory cortex (Gaab et al., 2006). However, it should be noted that the goal of our study was not to provide a comprehensive quantitative comparisons between sequences, either in terms of how different acoustic parameters may be affected or, particularly, possible differences in their statistical power, as has been reported in some studies (Adank, 2012; Gaab et al., 2006; Hall et al., 1999).

Likewise, the activation patterns obtained with the standard univariate categorical ANOVA were very similar to those yielded by a stimulus-based multivariate ICA. This increases our confidence that the findings are not driven by a few high-leverage distinct stimuli in each category. Moreover, the distribution of the ICA stimulus-specific coefficients supports the hypothesis that the activation patterns represent the acoustic processing of the stimuli, rather than their potential categorization performed (implicitly) by participants. Indeed, the singing stimuli whose coefficients were closest to instrumental music (i.e., most positive) were chorales, whereas those with most negative values (i.e., most similar to those from speech) included amateur singing, lullabies, and a melodic Torah reading. These findings suggest the presence of a gradient from speech to music, which may be dependent on the clarity of the speech in the stimulus, irrespective of comprehension. The distribution of clusters responding preferentially to one or more of the different stimulus categories, as show in Figure 1b, aligns with the model proposed by Peretz and Coltheart (2003), suggesting that numerous discrete modules are involved in music and language processing, some of which overlap, while others appear independent.

Our paradigm was also designed to minimize other potential confounding effects, such as stimulus expectation, by equalizing the number of stimuli in each of the three categories and all first-order transition probabilities, as well as counterbalancing, across subjects, the specific order of stimuli within and between runs. While there was no explicit task for the participants to perform, we cannot exclude the possibility that some of the participants performed some sort of stimulus categorization (although this was not reported in the debriefing following the experiment). Nonetheless, we believe our findings are unlikely to be purely the result of such putative cognitive task, as mentioned above. Moreover, a recent meta-analysis of examining the role of attention on processing of auditory stimuli, including voice, observed that no additional areas in auditory cortex were recruited in active, compared with passive, listening conditions (Alho, Rinne, Herron, and Woods, 2014).

5 | CONCLUSIONS

Different regions in the temporal lobe responded preferentially to vocal and musical stimuli. These included the superior temporal sulcus and gyrus for the former, and the planum polare and temporale for the latter. Consistent with its having both vocal and musical properties, singing recruited all these areas. Importantly, the results were obtained with a large and varied set of stimuli, as well as different acquisition sequences and analysis approaches. Taken together, these findings provide further support for a hierarchical processing of complex social acoustic stimuli along the temporal lobes, similar to what has been reported for the visual modality.

ACKNOWLEDGMENTS

We are grateful to Mike Ferreira and Ilana Leppert for invaluable help implementing the multiband and sparse sequences. This study was supported by grants from the Natural Science and Engineering Research Council of Canada (NSERC, 262439-2009, 2017-05832) and the Canadian Institutes of Health Research (CIHR, MOP-130516, MOP-97967) to JLA. JCW was funded by an NSERC-Create Auditory Cognitive Neuroscience Graduate Fellowship, a CRBLM Graduate Student Stipend, a McGill's Integrated Program in Neuroscience Returning Student Award, and an NSERC - Canadian Graduate Scholarship – Masters fellowship.

CONFLICT OF INTEREST

Authors have no conflicts of interest to declare.

ORCID

Jocelyne C. Whitehead https://orcid.org/0000-0002-8002-113X

REFERENCES

Adank, P. (2012). Design choices in imaging speech comprehension: An activation likelihood estimation (ALE) meta-analysis. *NeuroImage*, *8*, 360–369. https://doi.org/10.1016/j.neuroimage.2012.07.027

Alho, K., Rinne, T., Herron, T. J., & Woods, D. L. (2014). Stimulus-dependent activations and attention-related modulations in the auditory cortex: A meta-analysis of fMRI studies. *Hearing Research*, *307*, 29–41. https://doi.org/10.1016/j.heares.2013.08.001

Angulo-Perkins, A., Aubé, W., Peretz, I., Barrios, F. A., Armony, J. L., & Concha, L. (2014). Music listening engages specific cortical regions within the temporal lobes: Differences between musicians and nonmusicians. *Cortex*, *59*, 126–137. https://doi.org/10.1016/j.cortex.2014.07.013

Aubé, W., Angulo-Perkins, A., Peretz, I., Concha, L., & Armony, J. L. (2015). Fear across the senses: Brain responses to music, vocalizations and facial expressions. *Social Cognitive and Affective Neuroscience*, *10*, 399–407.

Armony, J. L., Aubé, W., Angulo-Perkins, A., Peretz, I., & Concha, L. (2015). The specificity of neural responses to music and their relation to voice processing: An fMRI-adaptation study. *Neuroscience Letters*, *593*, 35–39. https://doi.org/10.1016/j.neulet.2015.03.011

Belin, P., & Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport*, *14*, 2105–2109. https://doi.org/10.1097/00001756-200311140-00019

Belin, P., Zatorre, R. J., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Brain Research. Cognitive Brain Research*, *13*(1), 17–26. https://doi.org/10.1016/S0926-6410(01)00084-2

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*, 309–312. https://doi.org/10.1038/35002078

Besson, M., & Schön, D. (2001). Comparisons between language and music. *Annals of the New York Academy of Sciences*, *930*(1), 232–258. https://doi.org/10.1111/j.1749-6632.2001.tb05736.x

Bilalić, M. (2016). Revisiting the role of the fusiform face area in expertise. *Journal of Cognitive Neuroscience*, *9*, 1345–1357. https://doi.org/10.1162/jocn_a_00974

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot International*, *5*, 341–345.

Brandman, T., & Yovel, G. (2016). Bodies are represented as wholes rather than their sum of parts in the occipital-temporal cortex. *Cerebral Cortex*, *26*, 530–543. https://doi.org/10.1093/cercor/bhu205

Brown, S. (2000). The "musilanguage" model of music evolution. In N. L. Wallin, B. Merker, & S. Brown (Eds.), *The origins of music* (pp. 271–300). Cambridge, MA: MIT Press.

Callan, D. E., Tsytsarev, V., Hanawaka, T., Callan, A. M., Katsuhara, M., Fuuyama, H., & Turner, R. (2006). Song and speech: Brain regions involved with perception and cover production. *NeuroTmage*, *31*, 1327–1342. https://doi.org/10.1016/j.neuroimage.2006.01.036

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *The Journal of Neuroscience*, *23*, 3423–3431.

Dietterich, T. G., & Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, *2*, 263–286. https://doi.org/10.1613/jair.105

Ewender, T., Hoffmann, S., & Pfister, B. (2009). Nearly perfect detection of continuous f_0 contour and frame classification for TTS synthesis. In *Proceedings of Interspeech 2009* (pp. 100–103), Brighton, UK.

Fecteau, S., Armony, J. L., Joanette, Y., & Belin, P. (2004). Is voice processing species-specific in human auditory cortex? An fMRI study. *NeuroImage*, *23*, 840–848. https://doi.org/10.1016/j.neuroimage.2004.09.019

Fedorenko, E., McDermott, J. H., Norman-Haignere, S., & Kanwisher, N. (2012). Sensitivity to musical structure in the human brain. *Journal of Neurophysiology*, *108*, 3289–3300. https://doi.org/10.1152/jn.00209.2012

Fodor, J. A. (1983). *The modularity of mind: An essay of faculty psychology*. Cambridge, MA: MIT Press.

Gaab, N., Gabrieli, J. D. E., & Glover, G. H. (2006). Assessing the influence of scanner background noise on auditory processing. I. An fMRI study comparing three experimental designs with varying degrees of scanner noise. *Human Brian Mapping*, *28*, 703–720. https://doi.org/10.1002/hbm.20298

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds brain areas involved in face recognition. *Nature Neuroscience*, *3*, 191–197. https://doi.org/10.1038/72140

Glasberg, B. R., & Moore, B. C. J. (2002). A model of loudness applicable to time-varying sounds. *Journal of the Audio Engineering Society*, *50*, 331–342.

Hall, A. J., Brown, T. A., Grahn, J. A., Gati, J. S., Nixon, P. L., Hughes, S. M., … Lomber, S. G. (2014). There's more than one way to scan a cat: Imaging cat auditory cortex with high-field fMRI using continuous or sparse sampling. *The Journal of Neuroscience Methods*, *224*, 96–106. https://doi.org/10.1016/j.jneumeth.2013.12.012

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliot, M. R., … Bowtell, R. W. (1999). "Sparse" temporal sampling in auditory fMRI. *Human Brain Mapping*, *7*, 213–223. https://doi.org/10.1002/(SICI)1097-0193(1999)7:3<213::AID-HBM5>3.0.CO;2-N

Jackendoff, R. (2009). Parallels and nonparallels between language and music. *Music Perception*, *26*, 195–204. https://doi.org/10.1525/mp.2009.26.3.195

Juslin, P. N. (2001). Communication emotion in music performance: A review and a theoretical framework. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and emotion: Theory and research* (pp. 309–397). New York, NY: Oxford University Press.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, *17*, 4302–4311. https://doi.org/10.3410/f.717989828.793472998

Kriegstein, K. V., & Giraud, A. L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *NeuroImage*, *22*, 948–955. https://doi.org/10.1016/j.neuroimage.2004.02.020

Lai, J., Pancaroglu, R., Oruc, I., Barton, J. J., & Davies-Thomson, J. (2014). Neuroanatomic correlates of the feature-salience hierarchy in face processing: An fMRI-adaptation study. *Neuropsychologia*, *53*, 274–283. https://doi.org/10.1016/j.neuropsychologia.2013.10.016

Lartillot, O., Eerola, T., Toiviainen, P., & Fornari, J. (2008) Multi-feature modeling of pulse clarity: Design, validation and optimization. In J. P. Bello, E. Chew, & D. Turnbull (Eds.), *Proceedings of the 9th international conference on music information retrieval* (pp. 521–526). Philadelphia, PA: Drexel University.

Lartillot, O., Toiviainen, P., & Eeorla, T. (2008). A matlab toolbox for music information retrieval. In C. Preisach, et al. (Eds.), *Data analysis, machine learning and applications* (pp. 261–268). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-78246-9

Leaver, A. M., & Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: Effects of acoustic features and auditory objective category. *The Journal of Neuroscience*, *30*, 7604–7612. https://doi.org/10.1523/JNEUROSCI.0296-10.2010

Livingstone, S. R., Peck, K., & Russo, F. A. (2013). Acoustic differences in the speaking and singing voice. *The Journal of the Acoustical Society of America*, *19*(1), 035080. https://doi.org/10.1121/1.4806630

Manan, H. A., Yusoff, A. N., Franz, E. A., & Mukari, S. Z.-M. S. (2013). The effects of background noise on brain activity using speech stimuli on healthy young adults. *Neurology*, *Psychiatry and Brain Research*, *19*, 207–215. https://doi.org/10.1016/j.npbr.2013.09.002

Murphy, D. R., Craik, F. I., Li, K. Z., & Schneider, B. A. (2000). Comparing the effects of aging and background noise of short-term memory performance. *Psychology and Aging*, *15*, 323–334. https://doi.org/10.1037/0882-7974.15.2.323

Nebel, K., Stude, P., Wiese, H., Müller, B., de Greiff, A., Forsting, M., … Keidel, M. (2005). Sparse imaging and continuous event-related fMRI in the visual domain: A systemic comparison. *Human Brain Mapping*, *24*, 130–143. https://doi.org/10.1002/hbm.20075

Nichols, T., Brett, M., Andersson, J., Wager, T., & Poline, J. B. (2005). Valid conjunction inference with the minimum statistic. *NeuroImage*, *25*, 653–660. https://doi.org/10.1016/j.neuroimage.2004.12.005

Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, *88*, 1281–1296. https://doi.org/10.1016/j.neuron.2015.11.035

Ohnishi, T., Matsuda, H., Asada, T., Aruga, M., Hikarata, M., Nishikawa, M., … Imabayashi, E. (2001). Functional anatomy of musical perception in musicians. *Cerebral Cortex*, *11*, 754–760. https://doi.org/10.1093/cercor/11.8.754

Patterson, R. D., Uppenkamp, S., Johnsrude, I. S., & Griffiths, T. D. (2002). The processing of temporal pitch and melody information in auditory cortex. *Neuron*, *36*, 767–776. https://doi.org/10.1016/S0896-6273(02)01060-7

Peelle, J. E. (2014). Methodological challenges and solutions in auditory functional magnetic resonance imaging. *Frontiers in Neuroscience*, *8*, 1–13. https://doi.org/10.3389/fnins.2014.00253

Peretz, I., & Coltheart, M. (2003). Modularity of music processing. *Nature Neuroscience*, *6*, 688–691. https://doi.org/10.1038/nn1083

Pernet, C. A. (2015). The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage*, *119*, 164–174. McAleer, P, Latinus, M, Gorgolewski, K. J, Charest, I, Bestelmeyer, P. E, Belin, P, https://doi.org/10.1016/j.neuroimage.2015.06.050

Peretz, I., Vuvan, D., Lagrois, M. É., & Armony, J. L. (2015). Neural overlap in processing music and speech. *Philosophical Transactions of the Royal Society B*, *370*, 20140090. https://doi.org/10.1098/rstb.2014.0090

Pinker, S. (1997). *How the mind works*. New York, NY: W.W. Norton.

Rabbit, P. M. A. (1968). Channel-capacity, intelligibility and immediate memory. *The Quarterly Journal of Experimental Psychology*, *20*, 21–248. https://doi.org/10.1080/14640746808400158

Rogalsky, C., Rong, F., Saberi, K., & Hickok, G. (2011). Functional anatomy of language and music perception: Temporal and structural factors investigated using functional magnetic resonance imaging. *The Journal of Neuroscience*, *31*, 3843–3852. https://doi.org/10.1523/JNEUROSCI.4515-10.2011

Saitou, T., Goto, M., Unoki, M., & Akagi, M. (2007). Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 215–218). New Paltz, NY: IEEE. https://doi.org/10.1109/ASPAA.2007.4393001

Saitou, T., Tsuji, N., Unoki, M., & Akagi, M. (2004). Analysis of acoustic features affecting "singing-ness" and its applications to singing-voice synthesis from speaking-voice. In *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP2004)* (pp. 1929–1932). Jeju Island, Korea.

Schirmer, A., Fox, P. M., & Grandjean, D. (2012). On the spatial organization of sound processing the temporal lobe: A meta-analysis. *NeuroImage*, *63*(1), 137–147. https://doi.org/10.1016/j.neuroimage.2012.06.025

Schön, D., Gordon, R., Campagne, A., Magne, C., Astésano, C., Anton, J. L., & Besson, M. (2010). Similar cerebral networks in language, music and song perception. *NeuroImage*, *51*, 450–461. https://doi.org/10.1016/j.neuroimage.2010.02.023

Schwarzbauer, C., Davis, M. H., Rodd, J. M., & Johnsrude, I. (2006). Interleaved silent steady state (ISSS) imaging: A new sparse imaging method applied to auditory fMRI. *NeuroImage*, *29*, 774–782. https://doi.org/10.1016/j.neuroimage.2005.08.025

Setsompop, K., Gagoski, B. A., Polimeni, J. R., Witzel, T., Wedeen, V. J., & Wald, L. L. (2012). Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty. *Magnetic Resonance in Medicine*, *67*, 1210–1224. https://doi.org/10.1002/mrm.23097

Sperber, D., & Hirschfield, L. A. (2004). The cognitive foundations of cultural stability and diversity. *Trends in Cognitive Science*, *8*, 40–46. https://doi.org/10.1016/j.tics.2003.11.002

Thaut, M. H., Trimarchi, P. D., & Parson, L. M. (2014). Human brain basis of musical rhythm perception: Common and distinct neural substrates for meter, tempo, and pattern. *Brain Sciences*, *4*, 428–452. https://doi.org/10.3390/brainsci4020428

Vieillard, S., Peretz, I., Gosselin, N., Khalfa, S., Gagnon, L., & Bouchard, B. (2008). Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognitive Emotion*, *22*, 720–752. https://doi.org/10.1080/02699930701503567

Woods, D. L., Stecker, G. C., Rinne, T., Herron, T. J., Cate, A. D., Yund, E. W., … Kang, X. (2009). Functional maps of human auditory cortex: Effects of acoustic features and attention. *PLoS One*, *4*, e5183. https://doi.org/10.1371/journal.pone.0005183

Zatorre, R. J., & Baum, S. R. (2012). Musical melody and speech intonation: Singing a different tune. *PLoS Biology*, *10*, e1001372. https://doi.org/10.1371/journal.pbio.1001372

How to cite this article: Whitehead JC, Armony JL. Singing in the brain: Neural representation of music and voice as revealed by fMRI. *Hum Brain Mapp*. 2018;39:4913–4924. https://doi.org/10.1002/hbm.24333